Automated Machine Learning (AutoML)

SLIDES BY:

LYDIA ZHENG and JIANNAN WANG

https://www.cs.sfu.ca/~inwang/

Motivation

- 1. Machine learning is very successful
- 2. To build a traditional ML pipeline:
 - Domain experts with longstanding experience
 - Specialized data preprocessing
 - > Domain-driven meaningful feature engineering
 - Picking right models
 - Hyper-parameter tuning
 - **>**

AutoML Vision

For Non-Experts

AutoML allows non-experts to make use of machine learning models and techniques without requiring to become an expert in this field first

https://en.wikipedia.org/wiki/Automated_machine_learning

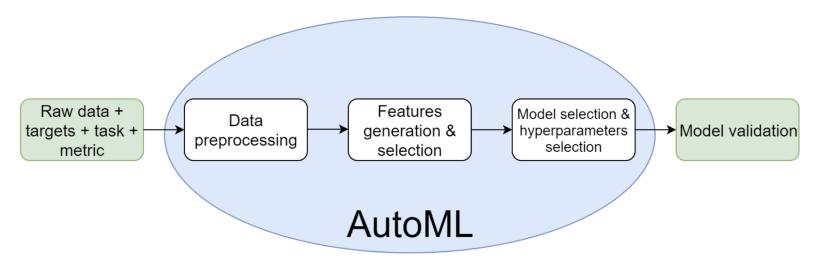
For Data Scientists

AutoML aims to augment, rather than automate, the work and work practices of heterogeneous teams that work in data science.

Wang, Dakuo, et al. "Human-Al Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated Al." Proceedings of the ACM on Human-Computer Interaction 3.CSCW (2019): 1-24.

What is AutoML?

Automate the process of applying machine learning to realworld problems



H20 Driverless Al Demo

https://www.youtube.com/watch?v=ZqCoFp3-rGc

EXPERIMEN	H20.ai (SN35). Current Use IT SETUP		ASSISTANT	STATUS: COMPLETE		TRAINING SETTINGS	EXPERT SETTINGS
DISPLAY NAME		DATASET		DEPLOY (LOCAL & CLOUD)			
CC Default Settings		CreditCard-train.csv		INTERPRET THIS MODEL		(5) (4) (6) (6 AUC
			TEST DATASET	DIAGNOSE MODEL ON NEW DATASET		(3)(4)(b) (AUC)
24K	25 0	PPED COLUMNS VALIDATION		SCORE ON ANOTHER DATASET		ACCURACY TIME INTO	ERPRETABILITY SCORER
24N	20 0			TRANSFORM ANOTHER DATASET			
dofault	bayment next	FOLD COLUMN		DOWNLOAD PREDICTIONS *		CLASSIFICATION REPRODUCIBLE	GPUS DISABLED
ueluult	Juyinent next			DOWNLOAD PYTHON SCORING PIPELINE			
WEIGHT COLUMN		TIME COLUMN		DOWNLOAD MOJO SCORING PIPELINE		CPU / MEMORY	Notifications Log Trace
		[OFF]		DOWNLOAD EXPERIMENT SUMMARY			
TYPE bool	23999	unique 2	TARGET FREQ 5369	DOWNLOS TO PEPONT		мем	
ITERATION	DATA - VALIDATION	1		VARIABLE IMPORTANCE		ROC P-R LIFT GA	UNS K-S SUMMARY
0 2 4 6 8 10 12 14 16 18			~~~	-	Experiment: CC Default Settings (b557b272-0649-11ea-8c75-0242ac110002) Version: 1.8.0. 2019-11-13 19:16		
			持以	639	Version: Lab., 2019-11-13 2019-15 Settings: 5/4/6, seed-709207965, GPUs disobled Train data: CreditCard-train.csv (23999, 25) Volldotion data: V/A		
-				8,UMT,BAL	0.12		singny 22 222% toront cines)
		+ + + 1 1		EEXX-XXII	839	System specs: Docker/Linux, 61 GB, 16 CPU cores, 0	VO GPU
		7700		0_AGE 4_BILL_AMT4	0.09	Max memory usage: 0.532 GB, 0 GB GPU Recipe: AutoOL (17 iterations, 4 individuals) Validation scheme: stratified, 1 internal holdou	
		//XY				Validation scheme: stratified, 1 internal haldou Feature engineering: 295 features scored (24 s	
	(Mode	el: [LIGHTGBM])		2.SAY AMTS	8.89	Timing: Data preparation: 4.89 secs Shift/Leakage detection: 6.01 secs	
				2.BILL_AMT2	0.07	Model and feature tuning: 63.38 secs (14 of 16	models trained)

Outline

Auto Feature Selection (Lecture 6)

Auto Hyperparameter Tuning (Lecture 6)

Auto Feature Generation (This Lecture)

Neural Architecture Search (This Lecture)

Auto Feature Generation

Motivation

- The model performance is heavily dependent on quality of features in dataset
- It's time-consuming for domain experts to generate enough useful features



Feature Generation

- Unary operators (applied on a single feature)
 - Discretize or normalize numerical features
 - Apply rule-based expansions of dates
 - Mathematical operators (e.g., Log Function)
- Higher-order operators (applied on 2+ features)
 - Basic arithmetic operations (e.g., +, -, ×, ÷)
 - Group-by Aggregation (e.g., GroupByThenMax, GroupByThenMin)

Featuretools



An open source library for performing automated feature engineering

Design to fast-forward feature generation across multi-relational tables

Concepts

- Entity is the relational tables
- An EntitySet is a collection of entities and the relationships between them
- Feature Primitives
 - Unary Operator: transformation (e.g., MONTH)
 - High-order Operator: Group-by Aggregation (e.g., GroupByThenSUM)

Entity sets

Customer

Product

Customer_id	Birthdate	MONTH(Birthdate)	SUM(Product.Price)
1	1995-09-28	9	\$500
2	1980-01-01	1	
3	1999-02-02	2	

GroupBy ThenSUM:

Product_id	Customer_id Name		Price
1	1	Banana	\$100
2	1	Banana	\$100
3	1	Orange	\$300
4	2	Apple	\$50
			•••

Unary Operator:

MONTH -

Feature Primitives

Outline

Auto Feature Selection (Lecture 5)

Auto Hyperparameter Tuning (Lecture 5)

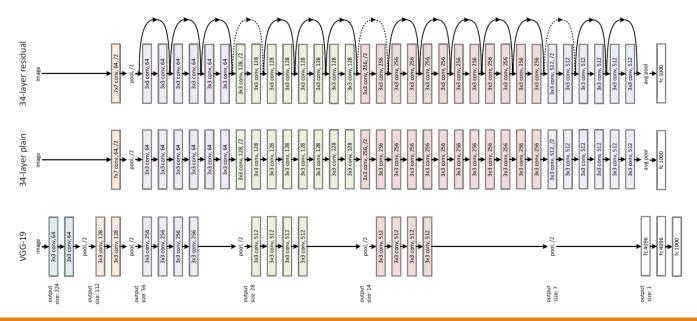
Auto Feature Generation (This Lecture)

Neural Architecture Search (This Lecture)

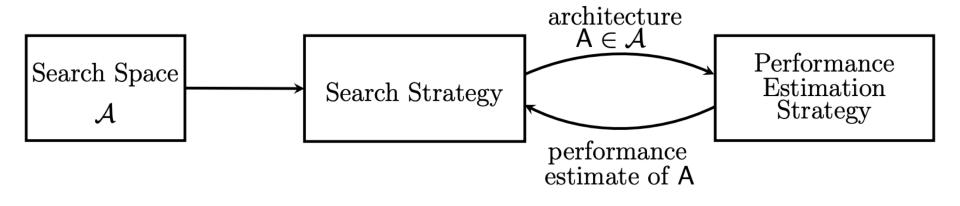
Neural Architecture Search (NAS)

Motivation

How can someone come out with such an architecture?

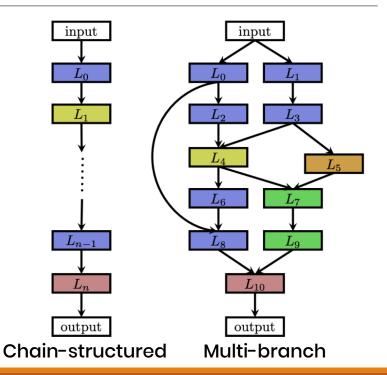


Neural Architecture Search: Big Picture



Search Space

- Define which neural architectures a NAS approach might discover in principle
- ♦ May have human bias → prevent finding novel architectural building blocks



Search Strategy

Basic Idea

> Explore search space (often exponentially large or even unbounded)

Methods

- Random Search
- Evolutionary Methods [Angeline et al., 1994]
 Bayesian Optimization [Bergstra et al., 2013]
- Reinforcement Learning [Baker et al., 2017]

Performance Estimation Strategy

Basic Idea

The process of estimating predictive performance

Methods

- Simplest option: perform a training and validation of the architecture on data
- Initialize weights of novel architecture based on weights of other architectures have been trained before
- Using learning curve extrapolation [Swersky et al., 2014]
- **>**

Summary

What is AutoML and why we need it? How AutoML works?

- Auto Feature Selection (Lecture 5)
- Auto Hyperparameter Tuning (Lecture 5)
- Auto Feature Generation (This Lecture)
- Neural Architecture Search (This Lecture)